A Q-values Sharing Framework for Multiple Independent Q-learners

Extended Abstract

Changxi Zhu
South China University of Technology
cxzhu.cn@gmail.com

Ho-fung Leung
The Chinese University of Hong Kong
lhf@cuhk.edu.hk

Shuyue Hu
The Chinese University of Hong Kong
syhu@cse.cuhk.edu.hk

Yi Cai
South China University of Technology
ycai@scut.edu.cn

ABSTRACT

By using a multiagent reinforcement learning (MARL) framework, cooperative agents can communicate with one another to accelerate the joint learning. In the *teacher-student* paradigm applied in MARL, a more experienced agent (advisor) can advise another agent (advisee) which action to take in a state. However, when agents need to cooperate with one another, the advisee may fail to cooperate well with others since their policies may have changed. It requires a long period for an advisee to learn the same best actions as an advisor has learned, especially when the amount of advice is limited. We propose a *partaker-sharer* advising framework (PSAF) for independent Q-learners with limited communication in cooperative MARL. In PSAF, the overall learning process is shown to accelerate by multiple independent Q-learners' sharing their maximum Q-values with one another at every time step. We perform experiments in the Predator-Prey domain and HFO game. The results show that our approach significantly outperforms existing advising methods.

KEYWORDS

multiagent learning; Q-learning; reinforcement learning

ACM Reference Format:
Changxi Zhu, Ho-fung Leung, Shuyue Hu, and Yi Cai. 2019. A Q-values Sharing Framework for Multiple Independent Q-learners. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement Learning (RL) [6] is successfully employed in many practical applications, such as robotics [5]. It is important to accelerate the learning for some complex domains, especially when the computing resource is limited. One notable approach is the *teacher-student* framework [8]. In this framework, a well-learned teacher agent advises its optimal action to a student agent in a state so that the student act optimally as the teacher. It is still an open question how agents benefit from the help of other agents when they are learning at the same time.

In fact, during learning process, each agent may have unique experiences or local knowledge of how to perform effectively in the task. By using a multiagent reinforcement learning (MARL) framework, agents can share their experiences, and learn from one another. Da Silva proposes a multiagent advising framework, where agents can advise one another while learning together [3]. An agent (advisee) can accelerate the learning by asking for action advices from a more experienced agent (advisor) in a state. Since we consider that agents independently learn to cooperate with others, they choose their individual actions, such that the resulting joint action is optimal. However, the advisee may fail to cooperate well with others even by following the suggested (generally sub-optimal) action, as the policies of all agents are ever changing.

In this paper, we consider a setting that the communication among agents are limited (e.g. communication cost). This setting is essential for some realistic problems, where agents need to take a lot of time to communicate because of the distance. In the case of cooperative agents, the advising strategy that uses actions as advices may not be good enough. In our work, each agent is independently learning its Q-function. Learning a policy means to estimate better Q-value for each action in every state. Intuitively, an agent (partaker) can ask for Q-values from a more experienced agent (sharer) in a state. After updating the requested Q-values, the partaker is more likely to perform effectively as the sharer in the state. We present a *partaker-sharer* advising framework (PSAF) for cooperative agents under limited communication. An agent can play the role of a partaker or a sharer in different sharing processes. There are two numeric budgets of each agent respectively, for requesting and providing Q-values. At each time step, if an agent does not visit the current state many times, it can take the role of partaker and asks for Q-values. The more times an agent updates its Q-value, the higher confidence it has in that Q-value. Only when the agent has higher confidence in its maximum Q-values than the partaker, can it take the role of sharer and provide the Q-values.

2 PRELIMINARIES

An Reinforcement Learning (RL) task is generally modelled as a Markov Decision Process (MDP) $\langle S, A, T, R \rangle$ [7] with a set S of states and a set A of available actions. The goal of an agent is to learn a policy $\pi : S \rightarrow A$ which maps states to actions in such a way that the expected cumulative discounted reward is maximized. Temporal difference (TD) RL algorithms such as Q-learning [9] and Sarsa [7] enable an agent to learn an action-value function, $Q(s, a)$, which is an estimate of the expected return that an agent takes action $a \in A$

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada. © 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

in state $s \in S$. A multiagent extension of MDP called Markov game [2] for N agents is defined as a tuple $\langle S, A_1, ..., A_N, T, R_1, ..., R_N \rangle$. In this paper, we focus on cooperative Multiagent Reinforcement Learning (MARL), where several RL agents jointly affect the environment and receive the same reward ($R_1 = R_2 = ... = R_N$).

3 Q-VALUES SHARING FRAMEWORK

We propose a *partaker-sharer* advising framework (PSAF) for multiple Q-learners under limited communication. In PSAF, all agents are cooperatively learning together in a shared environment. They can accelerate the overall learning process by sharing their maximum Q-values with one another. In a sharing process, a *partaker* is a role of an agent who asks for Q-values, and a *sharer* is a role of another agent who provides its maximum Q-values in the partaker's state. Each agent can play the role of a partaker or a sharer in different sharing processes. The maximum number of times that a partaker asks for Q-values and a sharer provides Q-values are b_{ask} and b_{give} respectively. Agents need to decide when to ask for Q-values and when to provide their maximum Q-values.

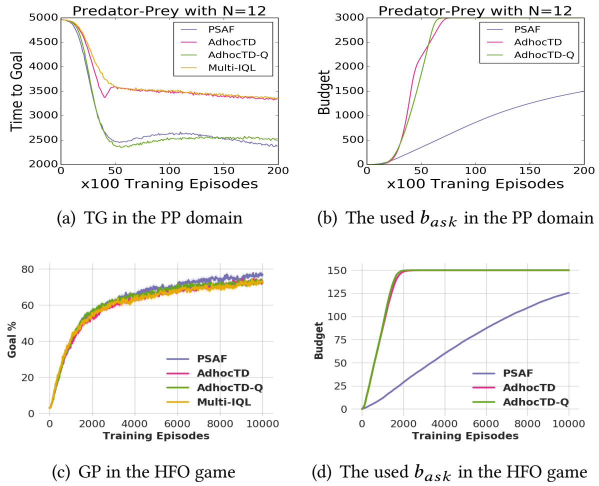
At each time step, agent i in state s_i may ask for Q-values as a partaker as long as the budget b^i_{ask} has not been used up. Intuitively, early in learning, an agent is more likely to ask for Q-values since many states have not been visited. As the agent visits a state more often, the estimated Q-values in the state becomes more reliable, and then the agent is less likely to ask for Q-values in the state. When agent i becomes a partaker in state s_i, it broadcasts to all other agents for requesting Q-values. If any sharer provides a Q-value in s_i, b^i_{ask} is decremented by 1. When partaker i receives several Q-values for a state-action pair, it randomly selects one of them. After that, partaker i replaces original Q-values with the selected Q-values for the corresponding state-action pair. Then the partaker executes its best action which corresponds to the maximum Q-value in the state. If agent i does not ask for Q-values or no Q-value is received, it uses ϵ-greedy as usual exploration strategy.

For an agent j that does not take the role of partaker, as long as budget b^j_{give} has not been used up, it may provide Q-values as a sharer. Notice that in PSAF, the sharer provides its maximum Q-value in a state rather than the whole Q-function at every time step. As agents are learning together in the same environment, their Q-values are generally not optimal. Agent j provides its maximum Q-value only when it has much higher confidence in the Q-value than partaker i. Intuitively, the more times a partaker updates its Q-value of a state-action pair, the higher confidence it has in that Q-value. When all Q-values of a sharer agent j in a state are the same, it does no matter which one is provided. If a sharer updates its maximum Q-value in a state many times, and the Q-value is much larger than other Q-values in this state, then the sharer has more confidence in its maximum Q-value. When agent j can provide its maximum Q-value, it takes the role of sharer and budget b^j_{give} is decremented by 1. Then sharer j sends its maximum Q-value (as well as the corresponding action) in s to partaker i. The partaker can accelerate the learning by updating its Q-values in s.

4 EXPERIMENTAL RESULTS

We compare PSAF with Multi-IQL, AdhocTD and AdhocTD-Q. In Multi-IQL, all agents are independent Q-learners and no sharing exists. In AdhocTD[3], each agent asks for and gives actions with probability P_{ask} and P_{give} respectively. In AdhocTD-Q, agents ask for Q-values and provide their maximum Q-values with the probability P_{ask} and P_{give} in AdhocTD respectively.

We evaluate PSAF, AdhocTD, AdhocTD-Q, and Multi-IQL in the Predator-Prey (PP) domain [1] and the Half Field Offense (HFO) game [4]. In all methods (except Multi-IQL), we set the value of $b_{ask} = b_{give}$. *Time to Goal* (TG) is the number of steps that predators take to catch the prey. Figure 1a shows that PSAF has a significantly lower TG than AdhocTD-Q after about 15,000 episodes. In Figure 1b, we can see that AdhocTD-Q completely spends all budgets after about 6,500 episodes. However, PSAF still has enough budgets that can be used, which leads to lower TG values. We use *Goal Percentage* (GP) for performance evaluation in the HFO game. GP is the percentage of episodes in which a goal is scored. Figure 1c shows that both PSAF has significantly higher GP than other methods after about 7,000 episodes. Before 4,000 episodes, AdhocTD-Q has higher GP than PSAF. During this interval, AdhocTD-Q quickly consumes all budget b_{give}, as shown in Figure 1d. All results show that PSAF consumes much less budget than other methods while it achieves a similar (even better) performance.



(a) TG in the PP domain

(b) The used b_{ask} in the PP domain

(c) GP in the HFO game

(d) The used b_{ask} in the HFO game

Figure 1: TG and used b_{ask} of PSAF, AdhocTD, AdhocTD-Q, and Multi-IQL in the PP domain with the size N=12 and the HFO game.

5 CONCLUSION

We propose a Q-values sharing framework PSAF for independent Q-learners in the cooperative MARL with limited communication. Our experiments show that Q-values sharing schemes, such as PSAF and AdhocTD-Q, are significantly better than actions advising schemes in two evaluation metrics TG and GP, yet PSAF spends much less budget than AdhocTD-Q.

ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Central Universities, SCUT (No. 2017ZD048, D2182480), the Science and Technology Planning Project of Guangdong Province (No. 2017B050506004), and the Science and Technology Program of Guangzhou (No. 201704030076,201802010027).

REFERENCES

[1] Tim Brys, Ann NowÃĪ, Daniel Kudenko, and Matthew Taylor. 2014. Combining Multiple Correlated Reward and Shaping Signals by Measuring Confidence. In *Proceedings of 28th AAAI Conference on Artificial Intelligence*. 1687–1693.

[2] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38 (2008), 156–172.

[3] Felipe Leno da Silva, Ruben Glatt, and Anna Helena Reali Costa. 2017. Simultaneously Learning and Advising in Multiagent Reinforcement Learning. In *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems*. 1100–1108.

[4] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, Eiichi Osawa, and Hitoshi Matsubara. 1997. RoboCup: A Challenge Problem for AI. *AI Magazine* 18 (1997), 73–85.

[5] Jens Kober, J. Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal Of Robotics Research* 32 (2013), 1238–1274.

[6] Michael L. Littman. 2015. Reinforcement learning improves behaviour from evaluative feedback. *Nature* 521 (2015), 445–451.

[7] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction* (1nd. ed.). MIT press, Cambridge, MA, USA.

[8] Lisa Torrey and Matthew E. Taylor. 2013. Teaching on a budget: agents advising agents in reinforcement learning. In *Proceedings of 12th the International Conference on Autonomous Agents and MultiAgent Systems*. 1053–1060.

[9] Christopher J.C.H. Watkins and Peter Dayan. 1992. Technical Note: Q-learning. *Machine Learning* 8 (1992), 279–292.